Neural Formatting for Spreadsheet Tables

Haoyu Dong, Jinyu Wang, Shi Han, Dongmei Zhang
Microsoft Research, Beijing 100080, China.
{hadong, wang.jinyu, shihan, dongmeiz}@microsoft.com

Abstract

Spreadsheets are popular and widely used for data presentation and management, where users create tables in various structures to organize and present data. Table formatting is an important yet tedious task for better exhibiting table structures and data relationships. However, without the aid of intelligent tools, manual formatting remains a tedious and time-consuming task.

In this paper, we propose CellGAN, a neural formatting model for learning and recommending formats of spreadsheet tables(Dong et al. 2020b). Based on a novel conditional generative adversarial network (cGAN) architecture, CellGAN learns table formatting from real-world spreadsheet tables in a self-supervised fashion without requiring human labeling. In CellGAN we devise two mechanisms, row/column-wise pooling and local refinement network, to address challenges from the spreadsheet domain. We evaluate the effectiveness of CellGAN against real-world datasets using both quantitative metrics and human perception studies. The results indicate remarkable performance gains over rule-based methods, graphical models or direct application of the state-of-the-art cGANs used in visual synthesis tasks. Neural Formatting is the first step towards auto-formatting for spreadsheet tables with promising results.

In this paper, we propose techniques for table formatting style transfer, i.e., to automatically format a target table according to the style of a reference table(Dong et al. 2020a). Considering the latent many-to-many mappings between table structures and formats, we propose CellNet, which is a novel end-to-end, multi-task model leveraging conditional Generative Adversarial Networks (cGANs) with three key components to (1) model and recognize table structures; (2) encode formatting styles; (3) learn and apply the latent mapping based on recognized table structure and encoded style, respectively. Moreover, we build up a spreadsheet table corpus containing 5,226 tables with high-quality formats and 784 tables with human-labeled structures. Our evaluation shows that CellNet is highly effective according to both quantitative metrics and human perception studies by comparing with heuristic-based and other learning-based methods.

	A	B	C	D	E	F	G
1	CHARACTERISTIC	Employed wage and salary workers			Non employed usual weekly earnings		
2							
3		Total	Percent		Total	Number	
4							
5				Not			Not
6			Union	union		Represented	represented
7			members	members		by union	by union
8	Total	129378	0.124	0.876	8960	4904	5056
9	Age						
10	16 to 34 years old	47981	0.157	0.843	3421	1342	2079
11	35 to 54 years old	59495	0.294	0.706	3427	1867	1560
12	55 years and over	21901	0.256	0.744	3112	1695	1417
13	Industry						
14	Private sector	108073	0.076	0.924	7429	3238	4191
15	Agriculture	1057	NA	NA	NA	NA	NA
16	Nonagriculture	107016	NA	NA	NA	NA	NA
17	Public sector	21305	0.0368	0.9632	1531	1666	865

(a) A table with default formats

	A	B	C	D	E	F	G
1		Employed wage and salary workers			Not employed usual weekly earnings		
2							
3			Percent			Number	
4	CHARACTERISTIC	Total			Total		
5				Not			Not
6			Union	union		Represented	represented
7			members	members		by union	by union
8	**Total**	**129,378**	**12.4%**	**87.6%**	**9,960**	**4,904**	**5,056**
9	**Age**						
10	16 to 34 years old	**47,981**	15.7%	84.3%	**3,421**	1,342	2,079
11	35 to 54 years old	**59,495**	29.4%	70.6%	**3,427**	1,867	1,560
12	55 years and over	**21,901**	25.6%	74.4%	**3,112**	1,695	1,417
13	**Industry**						
14	Private sector	**108,073**	7.6%	92.4%	**7,429**	3,238	4,191
15	Agriculture	**1,057**	NA	NA	**NA**	NA	NA
16	Nonagriculture	**107,016**	NA	NA	**NA**	NA	NA
17	Public sector	**21,305**	3.7%	96.3%	**2,531**	1,666	865

(b) A table with human-crafted formats

Figure 1: Comparison of an example spreadsheet table with deault formats and human-crafted formats.

References

Dong, H.; Yang, J.; Han, S.; and Zhang, D. 2020a. Learning formatting style transfer and structure extraction for spreadsheet tables with a hybrid neural network architecture. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2389–2396.

Dong, H. D.; Wang, J.; Fu, Z.; Han, S.; and Zhang, D. 2020b. Neural formatting for spreadsheet tables. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 305–314.